DOCUMENT RESUME

ED 185 056 TM 800 024

AUTHOR Irvin, Larry K: And Others
TITLE Evaluation of Retarded Student Achievement in Career Education Programs. Final Project Report, November, 1978.
INSTITUTION Oregon Univ., Eugene. Rehabilitation Research and Training Center in Mental Retardation.
SPONS AGENCY Bureau of Education for the Handicapped (DHEW/OE), Washington, D.C.
BUREAU NO 443CH60117
PUB DATE Nov 78
GRANT G008605412
NOTE 39p.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Adolescents: *Daily Living Skills: High Schools: Item Analysis: *Mild Mental Retardation: *Multiple Choice Tests: Objective Tests: Practical Arts: Questionnaires: *Situational Tests: Special Education: Test Reliability: Test Validity: Verbal Tests
IDENTIFIERS Social and Prevocational Information Battery

ABSTRACT

Three tests of daily living skills for retarded adolescents were developed and validated: making purchases, seeking jobs, and using banks. These three tests were adapted from three of the nine sections of the Social and Prevocational Information Battery (SPIB), a test made up mostly of true-false items. Each of the three tests were developed in three different formats: true-false (like the SPIB), three-choice multiple-choice, and two-choice multiple-choice. Two other instruments were also developed to cover the same content: a behavior rating questionnaire for the parents for their evaluation of the adolescents, and an applied performance criterion instrument consisting largely of actual or real-life types of activities and materials. Test statistics for the preliminary form of the test were obtained for 323 students. Revised forms were developed and the tests were administered to a new sample together with the behavior rating form and the applied performance test. (The results of this testing are reported and discussed). (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED185056

IM800 024

Evaluation of Retarded Student Achievement
in Career Education Programs

Bureau of Education for the
Handicapped (BEH)

Grant #G008605412
Project #443CH60117

Final Project Report; November, 1978

Larry K. Irvin, Andrew S. Halpern and
Janet T. Landman

2/A

TABLE OF CONTENTS

	Page
List of Tables	ii
List of Figures	iii
Introduction	
Statement of the Problem	1
Overall Design	6
Objectives	7
Project Procedures	7
Knowledge Tests	7
Development	7
Sample and Initial Piloting Procedures	8
Initial Data Analyses	9
Refinement Sample and Procedures	9
Refinement Data Analyses	10
Comparison of Reliabilities	13
Behavior Rating Instrument	14
Development	14
Sample and Procedures	15
Data Analyses	16
Criterion Validity Instrument	16
Development	16
Comparison of Validity of Alternate Predictor Formats	18
Project Results	19
Introduction	19
Final Psychometric Characteristics of Instruments	21
Reliability of Alternate Forms	23
Validity of Alternate Forms	24
Cross-Validation	27
Discussion	32
References	33

LIST OF TABLES

	Page
Table 1 - Overall Test Statistics	11
Table 2a - Descriptive Statistics All Three Domains Combined	21
Table 2b - Purchasing Domain	22
Table 2c - Job Search Domain	22
Table 2d - Banking Domain	23
Table 3 - Reliability Estimates	24
Table 4a - Correlation Matrix and Multiple Regressions All Three Domains Combined	25
Table 4b - Purchasing Domain	26
Table 4c - Job Search Domain	26
Table 4d - Banking Domain	27
Table 5a - Cross-Validation of Simple Regressions All Three Domains Combined	28
Table 5b - Purchasing Domain	28
Table 5c - Job Search Domain	29
Table 5d - Banking Domain	29
Table 6a - Cross-Validation of Multiple Regressions All Three Domains Combined	30
Table 6b - Purchasing Domain	30
Table 6c - Job Search Domain	31
Table 6d - Banking Domain	31

LIST OF FIGURES

	Page
Figure 1 Knowledge Test Items	12
Figure 2 Behavior Rating Form Item	15
Figure 3 Applied Performance Test Item	18

INTRODUCTION AND STATEMENT OF THE PROBLEM

For the past 15 years, work/study programs have been a major educational and rehabilitation resource for preparing retarded adolescents to assume their ultimate roles as vocationally adequate adults. More recently, such programs have been integrated into the broader context of career education efforts that "systematically coordinate all school, family and community components to facilitate each individual's potential for economic, social and personal fulfillment" (Brolin, 1974, p. 5). In addition to the achievement of academic skills, program participants are expected to develop a variety of adaptive life skills before graduation from formal secondary school programs. Career education models such as Project PRICE (Brolin, 1974) and Toward Competency (Oregon State Department of Education, 1974) emphasize the importance of education for all potential adult roles--citizen, parent, family member, student, paid worker, homemaker, retiree, and others. Typically, these approaches outline broad domains of consensually established primary curriculum areas and identify the specific behavioral competencies in each that are postulated as critical for successful adult functioning.

Although work/study and career education programs have become highly developed in their objectives, curricula, and methods of operation, little progress has been made historically in the development of related assessment instruments. Brolin (1975, p. 4) notes the need for development of "measures for evaluating and monitoring each student's competency attainment" in career education programs. It is also essential to develop measurement strategies to evaluate the impact of career education programs on eventual community and vocational adjustment (Proceedings of the Conference on Research Needs Related to Career Education for the Handicapped. 1975, p. 80).

Measurement of Knowledge

Another concern which requires attention involves the issue surrounding knowledge vs. actual performance as the focus of assessment. In recent years, there has been a strong plea from several sources to focus our test development efforts upon "criterion sampling" (Kulman, et.al., 1975; Lavisky, 1975; McClelland, 1973). The appeal of applied performance testing lies not only in the common sense desireability of measuring what really is of interest but also in the assumption that measurement of criterion behavior has much clearer relevance for instructional intervention than measurement of some surrogate for the criterion.

There are numerous instances, however, in which the evaluation of knowledge is more relevant than the evaluation of applied performance. Perusal of any of the currently well-developed career education models (e.g., Project PRICE, Toward Competency, The California Work/Study Handbook, etc.) illustrates that a large portion of the identified objectives presuppose or directly call for the attainment of knowledge. A sampling of such objectives across many domains includes making wise expenditures, selecting housing, discharging marriage responsibilities, utilizing effective child rearing practices, understanding citizenship rights and responsibilities and governmental systems, understanding recreation values, attaining self-awareness and satisfaction, knowing socially acceptable behaviors, anticipating consequences of behavior, knowing appropriate interview behavior. Although the successful attainment of many of these objectives cannot be claimed without indications of applied performance, the measurement of knowledge is still important for at least two reasons: (1) many of the desired performance competencies depend upon a foundation of knowledge if they are to be learned as strategies rather than as rote responses to narrowly circumscribed classroom stimuli; and (2) many of the desired performance competencies can only be displayed long after the student has terminated from school (e.g., discharging marriage responsibilities).

Development of a Test Battery Measuring Adaptive Life Skills

Given that there are almost no measures of applied performance presently available to be used in career education programs for the mentally retarded, productivity in any direction is a welcome addition to the field. For a variety of reasons, the authors have focused both past and present research efforts on the development of measurement strategies and instruments that assess level of knowledge in the adaptive life skill areas. With the assistance of teachers from work/study programs in Oregon, pupil competencies stressed in Oregon's work/study programs were identified and translated into long-range, post-school goals and short-term classroom objectives. Thus, items were selected in accordance with a domain sampling model, whereby each domain or long-term goal area to be tested was further specified hierarchically into 54 content areas and 180 sub-content areas (corresponding to short-term classroom objectives). The sub-content areas then served as a blueprint for generating test items for possible inclusion in the battery. Extensive field testing produced 277 items, either true/false or picture selection in format, which are orally administered to examinees in order to neutralize the impact of differential reading ability. The Social and Prevocational Information Battery (SPIB) consists of nine tests, each of which contains approximately 30 items that can be administered separately, requiring 10 to 15 minutes for completion (Halpern, Raffeld, Irvin, & Link, 1975b).

The battery provides an assessment strategy and instrumentation for evaluating student achievement of knowledge-based competencies in the areas of job search skills, job-related behavior, banking, budgeting, purchasing, home management, physical health care, personal hygiene, and functional signs (survival reading). Research evidence concerning both the reliability and validity of the SPIB is extremely encouraging (Halpern, Raffeld, Irvin, & Link, 1975a; Irvin & Halpern, 1977).

Focus of the Present Study

Even though the SPIB is very adequate on a number of important dimensions of measurement, it would be premature to adopt the SPIB model as a paradigm for the construction of all future instruments to be used in career education programs for the mentally retarded. At least three major questions remain either unanswered or only partially answered, the first two involving item format and the third dealing with broader method of measurement effects:

--Can an orally administered multiple-choice format be employed with equal or greater reliability than an orally administered true/false format?

--Assuming adequate reliability, which of these item formats provides the most valid assessment of student behavior?

--Will a behavior rating measure prove as valid in predicting performance as direct testing within either of the above types of item format?

Item format. One might be tempted to answer the first of these questions without hesitation, since there is a general consensus among test developers that multiple-choice tests are usually superior to true/false tests (Chase, 1974; Cronbach, 1950; Frisbie, 1973, 1974; Gronlund, 1968, 1971, Nunnally, 1964; Thorndike & Hagen, 1961). Four arguments are typically offered in support of this conclusion:

1. Multiple-choice formats tend to produce greater diagnostic precision, because they are less susceptible than true/false formats to providing examinees with the opportunity to give the right answer for the wrong reason. In other words, well-constructed multiple-choice formats can elicit examinee rationale for responses, which is difficult if not impossible for true/false formats to accomplish.

2. Multiple-choice tests can assess achievement of a greater variety of educational objectives. Chase (1974) notes that all levels of a taxonomy of educational objectives

can be sampled via multiple-choice items. Alternatively, true/false items tend to be more restricted in the kind of educational objective they can measure (Chase, 1974; Thorndike & Hagen, 1961).

3. As the number of alternative response possibililties increases, the probability of guessing a correct response decreases, and the probability of higher relilability increases (Ebel, 1969; Frisbie, 1973, 1974; Remmers, Gage & Rummel, 1965).

4. Multiple-choice tests tend to be free of response set effects which are particularly influential on true/false tests, at least with non-retarded populations (Cronbach, 1950; Gronlund, 1968).

Given the apparent strength of these arguments, one might question the choice of the SPIB format which consists primarily of orally administered true/false items. This decision was influenced by assumptions, knowledge, and experience relating to the reading and short-term memory abilities commonly exhibited by retarded adolescents in career education program.

Reading ability was not considered integral to knowledge of many social and prevocational competencies. Because reading skills are so variable within groups of retarded adolescents, the SPIB test administration was designed to eliminate reading ability as a determinant of performance except where reading skill is a critical part of a competency being learned, e.g., reading signs in the everyday environment.

Some degree of short-term retention ability is required for performance on orally presented test items. During SPIB development, deficits in short-term memory skills were assumed to be present to some extent in the population of retarded adolescents. The choice of an orally administered true/false and picture selection format derived directly from this assumption. Furthermore, there is evidence that low achieving students perform better on orally administered true/false achievement tests than on the same tests presented in a written form (Odom, & Miles, 1951). On the other hand, it was assumed that a verbally presented mulitple-choice item would be too difficult for retarded students since the stem and all alternatives of such an item must be remembered and analyzed, and the order of each alternative must also be remembered in order to choose correctly the corresponding symbol or number on an answer sheet.

The argument has now gone full circle, from the theoretical desireability of a multiple-choice format, to the presumed unfeasibility of an orally administered multiple-choice format for use with a retarded population. This presumption, however, has not been and yet should be evaluated, since it is possible to construct multiple-choice items with short stems and short alternatives, which will reduce the effects of limited auditory retention

on test performance. Given the potentially better psychometric properties of multiple-choice tests, as well as findings indicating the superiority of orally presented tests for use with other students described as "inferior" in achievement, we should not presume what retarded adolescents can and cannot do.

Reynolds (1976) conducted a pilot study to investigate the utility of multiple-choice test formats with mildly retarded adolescents. In that study, the Hygiene and Grooming Test of the SPIB was modified from a true/false format to two different multiple-choice formats--two-choice and three-choice. The resulting 36-item tests were orally administered to classroom groups of mildly retarded adolescents. The sample consisted of 100 subjects with a mean IQ of 73.5 and a mean chronological age of 16.4 years.

Reynolds' (1976) results were promising in that it was demonstrated that retarded adolescents could respond appropriately to the multiple-choice test used in the study. However, these findings were equivocal in several respects: (1) the tests were quite easy, which provided no information about the adequacy of assessment of difficult concepts, (means for the two-choice and three-choice forms were 86% and 82% correct, respectively); (2) item validity was moderate (the average inter-item correlation was $r_{ij} = .07$ for both the two-and three-choice forms, and (3) internal consistency reliability was also moderate (coefficient alpha reliability was $r_\alpha = .73$ and $r_\alpha = .72$ for the two-and three-choice forms, respectively) and was based on a sample of only 100 subjects.

The intent of the item format investigation within the current study was to further establish the utility of multiple-choice formats with retarded adolescents in such a way that: (1) the number of subjects is large; (2) assessment in more than one different content areas is accomplished; (3) optimum mean difficulties for three-choice multiple-choice tests, are obtained, i.e., means that approximate the point that is halfway between chance and maximum scores; and (4) optimum variability, item validity and test internal consistency reliability indices are obtained.

Criterion validity. The validity of any test is surely its most difficult characteristic to establish. Without such information, however, it is impossible to ascertain the real utility of a test. Because of its ultimate importance, the major purpose of this project has been to investigate the validity of the several item and instrument formats that have been discussed.

Although studies conducted with the SPIB provide us with some information on the predictive and concurrent validity of this type of instrument with a retarded population (Halpern, et al., 1975a; Irvin and Halpern, 1977; Irvin, Halpern, & Reynolds, 1977), these investigations can be improved upon in at least two ways.

First, the predictors used in earlier studies were limited to direct tests of knowledge. However, most measures of adaptive life skills are derived from rating scales that are administered to third-party informants who are presumed to be well acquainted with the person being evaluated. To a much lesser extent, direct testing has been used as an approach for measuring adaptive behavior. In a recent comprehensive review of the field, Meyers, Nihira, and Zetlin (1978, in press) identify only two fully developed instruments of this latter sort: the Social and Prevocational Information Battery (SPIB), and one subscale of the Balthazar Scales of Adaptive Behavior. Such reviews of the extant literature clearly lead to the conclusion that current practice is largely dependent on the behavior rating approach. Whether or not it will remain the backbone of measurement in career education in the future, however, is open to continuing depate and empirical investigation. Another important objective of this study, then, was to examine and compare the relative efficacy of a rating and a testing approach in predicting performance within career education domains for use with mildly retarded adolescents and young adults.

A second limitation of prior validity studies arises from the possible inadequacy of the criteria used: fairly broad rating-type measures provided by rehabilitation counselors and community residential facility programming personnel. A thorough evaluation of the validity of either the testing or rating approach requires quantification of the extent to which measured behavior provides an adequate index of the criterion behavior that is of ultimate concern. This criterion is usually defined as the demonstrated ability actually to perform a desirable behavior in a relevant context. At its most basic definitional level, then, measurement of the criterion of applied performance is accomplished when a desired behavior is measured in the context in which performance is expected. Unfortunately, however, life is simply too complex to serve either as a criterion or as a setting in which to measure all the precise behavioral criteria that are desired outcomes of instructional interventions or the desired behavioral dimensions for selection and classification decisions. One step removed from complete contextual relevance would be measurement of the desired behavior in a simulated context. A second step removed would be measurement of knowledge about the desired behavior (Lavisky, 1975). Whatever the method employed, measurement almost always requires a sampling of behavior. In the current project, the criterion established consists of a sampling both of applied behaviors in a simulated context and of knowledge about the desired behaviors within the relevant career education domains.

Overall Design of the Present Study

In order to implement the present study, it was necessary to administer to the same subjects analogous measures of knowledge, behavior rating, and applied performance. In addition to the original SPIB tests with their true/false format, two multiple-choice forms of knowledge tests were developed, consisting of orally administered items with either two or

three response options. A set of behavior rating forms and a set of applied performance tests were also developed, so that the final measurement package consisted of five instruments: three sets of knowledge tests, one set of behavior rating scales, and one set of applied performance tests. The study was designed to provide alternative measures of identical content, so that item format and method of measurement could be isolated as the relevant independent variables for the two major components of the study.

Objectives

The primary purpose of this project was to examine the reliability and validity of testing approaches (using true/false and multiple-choice item formats) and a rating approach for testing important student behaviors within career education programs for the mentally retarded. Six specific objectives were fulfilled.

Predictor Instrumentation

1. Development of preliminary three-option multiple-choice formats.
2. Refinement of revised multiple-choice forms and development of two-choice forms.
3. Comparison of reliability of alternate item formats (true/false, and two-option and three-option multiple-choice).
4. Development of behavior rating instrument.

Criterion Validity

5. Development of applied performance criterion instrument.
6. Comparison of the validity of alternate forms of item formats within knowledge tests, and of alternate methods of measurement (knowledge testing and behavior rating).

Procedures

Development of Preliminary Three-Option Multiple-Choice Formats

1. Test Development

Of the nine original true/false SPIB tests, three were selected for transformation into a multiple-choice format: Banking, Purchasing Habits,

and Job Search Skills. The following criteria were utilized in the selection process: (a) the presence of uniformly high reliabilities across all previously tested samples; (b) suitability for transformation into orally administered multiple-choice formats, as well as into behavior rating and applied performance instruments measuring skill behaviors in analogous content areas; and (c) representation across more than one of the five broad goal-domains that provided the content structure and rationale for the SPIB. Following test selection, multiple-choice versions of each test were constructed, using both the general SPIB content area specifications and existing SPIB items as sources for item content. As a result, approximately 88% of the multiple-choice items in each test were directly analogous in content to items in the corresponding SPIB true/false form. Additional items within each broad domain area were also constructed in order to ensure a sufficiently large item pool to withstand subsequent elimination of items on the basis of item analysis. Four alternatives were constructed for each item stem, in order to allow for the eventual empirical selection of three sound alternatives (a correct answer plus two plausible distractors). Item stems and alternatives were kept as brief as possible to minimize memory effects.

Prior to initial piloting with students, all items were reviewed by eight experienced high school career education teachers/program coordinators on dimensions of relevance, clarity, plausibility and difficulty. On the basis of these expert reviews, project staff made appropriate item revisions and deletions.

2. Sample and Procedures

Approximately 330 students from 30 classrooms in four states (California, Oregon, Texas and Washington) gave their informed consent to participate and were administered the preliminary versions of each multiple-choice test. In most cases, a local coordinator within participating school districts assumed liaison responsibility for obtaining local school participation, assuring compliance with procedures for protecting human subjects, and monitoring the flow of information and materials. In some instances, after such initial negotiations were completed by district liaison persons, all communication was carried on directly with participating teachers.

Through counterbalanced presentation of three of possible four options to each student, three-choice versions were field tested. With all participants, pilot testing was accomplished via classroom group administration of the three tests by teachers. Tests were typically administered in sessions of 25-45 minutes, thus requiring two or three sessions for all testing to be completed. Demographic information was also collected for participating subjects.

3. Data Analyses and Revision of Preliminary Forms

Initial pilot forms of the multiple-choice tests contained from 48-54 three-choice items. One content and two statistical criteria were established in order to guide the process of item elimination. Desireable statistical properties of items included the presence of difficulty levels not greater than .95 nor less than .30, and item/total test correlations above .15. At the same time, care was taken to maintain adequate representation of the various content and subcontent areas within each test.

As a result of test and item analyses guided by these three principles, the revised pilot forms of the three tests contained from 46-50 items, a reduction of 4-7% in total number of items per test. A distractor analysis was then accomplished with each of the retained items in order to achieve empirical selection of the two most plausible incorrect alternatives for retention in the revised pilot forms of each test. Finally, total test statistics, including inter-item correlations and coefficient alpha internal consistency reliabilities were calulated.

Mean difficulties of each of the three initial pilot tests were between 60% and 65%, thus approaching the optimal difficulty level of halfway between chance level of maximum score. Standard deviations ranged from 7.4 to 9.1 for sets of approximately 50 items in each test. The average interitem correlations ranged from .08 to .13 for the three tests, and coefficient alpha internal consistency estimates exceeded $r\alpha = .80$ for all tests.

Refinement of Revised Multiple-Choice Forms and Development of Two-Choice Forms

1. Sample and Procedures

Three hundred twenty-three students from thirty high school EMR classes in the same four states represented in the preliminary pilot sample were administered the revised item sets. This sample was almost identical to the preliminary sample, and demonstrated adequate representation on the variables of age, sex, grade distribution and ethnicity. The average age of participating students was 16.8 years, with 56% of the sample being male and 40% female (4% sex unspecified). Sixteen percent of the total sample were 9th graders, with 23%, 27%, and 25% of the sample composed of students from grades 10, 11, and 12 respectively. The sample consisted of 60% Anglo, 27% Black, 1% Oriental or American Indian, and 9% Spanish-surnamed students.

Although the principal purpose of this second phase of testing was to gather further information regarding multiple-choice item and test characteristics, an ancillary investigation was carried out to determine whether the orally administered multiple-choice format might elicit unwanted response styles in an EMR population, e.g., a systematic tendency to choose the first (or second or third) alternative when guessing. In order to test for this possibility, the items within each test were randomly divided into three equal groups of items: one group to be keyed a, another to be keyed b, and the third to be keyed c. A second form of each test was then created by rotating the position of the correct answer one place, leaving the content unchanged; and a third form consisted of the same items with the position of the correct answer rotated one more place. Using the previous class test results as a guide, project staff assigned one of the three test forms to each class, attempting to obtain three groups equivalent both in size and level of previous test performance. Group A originally contained 112 students with an average previous test score of 30.1; Group B consisted of 109 students with a mean prior test score of 30.7; and Group C, 115 students whose average score on the preliminary tests had been 30.3. As teachers in two of the Group B classrooms were unable to carry out testing, the final total number of students tested on the revised three-option multiple-choice tests was 323.

2. Data Analyses

The primary analyses performed were the examination of item difficulty levels and item/test correlations for each item within each test, in order to gain additional information for the refinement of the three-option multiple-choice tests and for the empirical development of two-choice tests. Using the item analysis criteria outlined in Objective 1 above, items with inadequate psychometric properties were eliminated.

Overall test statistics were also calcualted and taken into account when making decisions involving elimination of items. The following overall test statistics were achieved for each test following elimination of unsuitable items: average inter-item correlations of $r_{ij} = .08$ or greater, coefficient alpha internal consistency reliability estimates of .80 or greater, and overall test difficulties of 58 to 68% of total items correct. Standard deviations ranged from 6.8 to 10.6 for sets of 46-50 items per test. Table 1 shows the results of the development portion of this phase of the project.

Table 1

Overall Test Statistics*

	No. of Items	Mean	SD[a]	p[b]	r[c]
Purchasing Habits	50	32.29	8.70	.65	.88
Banking	50	30.45	9.80	.61	.90
Job Search Skills	46	27.42	7.90	.60	.85

*No. of Students = 323

[a]Standard deviation

[b]Average proportion of items correct

[c]Coefficient alpha internal consistency reliability

A third type of analysis performed consisted of the examination of test results in terms of the presence and magnitude of response style (specifically, position preference) elicited by the multiple-choice format. According to the position preference paradigm, if examinees systematically prefer one response (a, b, or c) over others, scores on items in which the correctly keyed alternative matches the preferred position should be characterized by the highest mean, the lowest variance, and the lowest internal consistency. Although the results of the position preference portion of this study did not conclusively demonstrate a preference for any single position, they did suggest a tendency for examinees to choose c and to avoid choosing a. This was concluded from the fact that in 7 of 18 comparisions, c subtests had the highest mean, the lowest variance, and the lowest internal consistency.

Based on these findings, a strategy for reducing the small but significant effects of position preference was carried out. Accordingly, the ratio of items keyed a, b, and c was set at 40% a, 30% b, and 30% c for the three-option version of each test included in the final reliability/validity field study. Items keyed a, b, and c were then scattered randomly throughout each test.

3. Development of Two-Option Multiple-Choice Forms

Subsequent to the development of psychometrically sound three-choice item sets following the initial pilot tryouts, two-option forms were developed empirically within each test. The criterion for selection of a two-choice item form consisted simply of degree of plausibility of distractors. Thus, the least plausible of the two distractors that had been administered in both pilot administrations was excluded from the two-alternative multiple-choice version of each item within each of the three tests. The resulting two-choice multiple-choice tests were exactly parallel in content to the three-choice tests, except that only one of the two incorrect alternatives from the three-choice items was included in the two-choice items. Finally, since the position preference portion of the study had addressed only the three-option format and since there seemed to be less reason to hypothesize such a position preference in the two-alternative version, the two-option format was designed to contain approximately equal numbers of a- and b-keyed items scattered randomly throughout each test. An example of an item parallel in content, developed within each of three knowledge testing item formats, is presented in Figure 1.

SPIB True/False Item

If something is sold at a discount, it means that the price has been raised. *Repeat item.*

2 - Choice Item

If something is sold at a discount, its price is

a. lowered.

b. raised. *Repeat item.*

3 - Choice Item

If something is sold at a discount, its price is

a. kept the same.

b. raised.

c. lowered. *Repeat item.*

Figure 1

Knowledge Test Items

Comparison of Reliabilities of Alternate Item Formats

1. Sample and Procedures

Additional new samples of students were sought for participation in the final knowledge testing phase of the project, resulting in the acquisition of an additional number of students from classrooms in Arizona, Minnesota, and Ohio to supplement the initial pilot groups from California, Oregon, Texas, and Washington. The mean age of this sample was almost identical to that of the earlier sample, 16.7 years, as was the sex distribution: 43% female, 54% male, and 3% sex unspecified. Representation of students at each grade level was somewhat different from that of the earlier sample, with 11%, 32%, 31%, and 22% of the total sample made up of students at grades 9, 10, 11, and 12 respectively. Students from several ethnic groups comprised the sample, with 67% of the total sample identified as Anglo, 22% Black, 3% either Asian American or American Indian and 5% Spanish-surnamed (3% of the students were unidentified according to ethnic category).

After consent had again been obtained from students and their parents, each student in the sample was administered two formats (true/false and/or two-choice and/or three-choice) of each of the three tests (Banking, Purchasing and Job Search Skills). In order to control for possible practice effects, the order of administering the two forms of each test was systematically counterbalanced. Care was also taken to assign approximately equal numbers of students to each of the three formats.

Project participants were assigned to one of two groups in order to permit concurrent administrations of knowledge and criterion tests. One group's knowledge testing period was scheduled for October through December of 1977, with the knowledge testing of the second group taking place from January through March of 1978. Within each of the two time frames, each student received Banking, Job Search, and Purchasing Habits tests within one format, followed in two or three weeks by the same three tests within a second item format. A total of 730 test booklets were returned from 425 students with 260 students having received the original true/false format of the Social and Prevocational Information Battery (SPIB), and 270 students having taken the two-option multiple-choice format, and 200 the three-choice form.

2. Data Analyses

In order to ascertain the impact of prior item revisions, item difficulty levels and item/test correlations were calculated for each item within each format of each test. Means, average difficulty levels, and standard deviations

were also determined for each form of each test. Results of these overall test analyses are summarized in Tables 2a to 2d on pp. 21 to 23 and are discussed in the Results section on p. 19. Table 3 on p. 24 presents the estimates of internal consistency reliability for each instrument used in the investigation.

Development of Behavior Rating Instrument

1. Instrument Development

Behavior rating instrument development was accomplished in much the same way as multiple-choice test development. That is, original SPIB content area and sub-content area specifications were used to guide development of items in this instrument. Whenever possible, an item from the knowledge tests was directly translated into an analogous behavior rating item. When this was not possible, new behavior rating items were constructed from the same content and sub-content areas that guided the development of the knowledge tests, i.e., Banking, Purchasing, and Job Search Skills.

Each item in the rating scale began with a behavioral description, such as "reads newspaper ads in order to look for a job," or "manages own checking account." Four response options were available, scaled along the dimension of independence:

--yes, with no help

--yes, with little help

--yes, with much help, and

--no

A card with all possible responses printed on it was provided to parents to facilitate their responses. Figure 2 shows an example of a behavior rating form item.

If the parent responded that a behavior had not been performed, a further attempt was made to determine whether the perceived absence of behavior was the result of lack of ability as distinguished from lack of motivation or opportunity. The informant was asked to indicate whether or not the person could perform the indicated behavior, with or without assistance.

SAY: "Please use the pink side of your card to answer this next question.

Number 21. Does _____ figure out the correct price
(Name of child)

of an item with a percentage discount?"

A. Yes, with no help. _____

Yes, with a little help. _____

Yes, with much help. _____

If parent has selected any of the A options, please skip to the next page.

B.1. No. _____

I don't know. _____

B.2. SAY: "Do you THINK _____ could figure
(Name of child)

out the correct price of an item with a percentage discount,

even though you don't know for sure?"

Yes, with no help. _____

Yes, with some help. _____

No. _____

Figure 2

Behavior Rating Form Item

2. Sample and Procedures

After knowledge testing was completed, a sub-sample of 147 students from 25 classrooms in 16 cities and 6 states were evaluated on the behavior rating form. Three criteria determined participation in this phase of the study:

--students actually completed all three knowledge tests in each of two formats;

--teachers were willing to administer the behavior rating interview; and

--parents were willing to participate in the behavior rating interview.

Using the roster of students from each class who had completed the knowledge tests, random numbers were assigned by the staff to the students' names. Teachers were asked to try to interview parents according to the order of these numbers. Thus, if a teacher had agreed to interview six parents, s/he was to try to interview the parents of those children number one through six. If for any reason these parents could not be interviewed, the teacher was to proceed to the next number on the list and request an interview with the parent of that student and to continue in this fashion until the agreed upon number of interviews was completed. The teachers returned the parent interview forms (behavior ratings of students) through the mail. All those who participated in the behavior rating phase of the study also engaged in the criterion validity portion.

3. Data Analyses

Analyses of the psychometric properties of the behavior rating instrument, such as those previously outlined in relation to the knowledge tests, were performed for this instrument and are presented in the Results section of this report on pp. 19 to 32. In addition, validity analyses were carried out, including the computation of intercorrelations of the behavior rating measure with all other instruments used, as well as multiple regression analyses. Through the multiple regression analyses, a predictor set that included both the knowledge and the behavior rating approaches was examined for its power to predict applied performance. Finally, cross-validation of the resulting multiple correlations was carried out.

Development of Applied Performance Criterion Validity Instrument

The criterion instrument developed is an individually administered applied performance test (APT) in each of the three content areas identified above. The Purchasing APT has 23 scorable items, and both the Job Search test and the Banking APT's contain 28 items. The measures consist largely of actual and simulated real-life types of activities and materials. Administration of the total APT required approximately one hour per subject.

Item response requirements included filling out job applicaton forms, checks, and other common forms; reading newspaper ads, grocery labels, and

menus; and counting money, making change, and figuring out unit prices. A smaller number of items involved open-ended requests for important knowledge in the domains being measured. Whenever possible, an item from the knowledge tests was directly translated into an analogous applied performance item. When this was not possible new applied performance items were constructed from the same content sub-content areas that guided development of the knowledge tests.

Many of the items in the applied performance tests were moderately complex, consisting of several component parts. If a respondent did not satisfactorally perform the whole task upon request, the component parts of the task were then presented separately. For example, if an examinee did not fill out a check for a given amount after being requested to do so, the examiner then deliverd instruction line by line, giving a clue if necessary, such as "the date goes on this line." Most of the items in the applied performance test were formatted in this way so as to provide "clues" or second chances to complete the item successfully. If an acceptable response was not elicited following the clue, the examiner provided the correct answer before moving on to the next item. Because of this procedure, most examinees were highly motivated and thoroughly enjoyed taking the test, even if their performance was very low.

The scoring system for most items involved a four-point scale. No points were awarded for completely unsuccessful performance, one point was counted for performance after a clue, two points for slightly incorrect performance with no clue, and three points for correct performance without clues. This characteristic of the instrument corresponds to the availability of discrete response options on the behavior rating scale, calling for parent evaluation of the degree of assistance required by the student in order to perform each task.

The applied performance test was pilot tested with three groups of mildly mentally retarded persons. The first two rounds of pilot testing involved a group of post-high school clients from two offices of the Oregon Association for Retarded Citizens who were reported to be experiencing difficulties in the areas of vocational and daily living skills. In the third round of piloting, participants were a group of high school EMR students enrolled in local work-study programs. Results of pilot testing were generally positive, and appropriate revisions were made in the tests on the basis of these results prior to administration to project subjects.

Figure 3 shows an example of an applied performance test item. Items in both the behavioral rating form and the applied performance test are somewhat more complex than their knowledge test counterparts, as can be seen by examining and comparing the analogous items presented in Figures 1, 2, and 3.

Task: Compute the percentage discount price

Materials: discount ads, paper & pencil

Directions for administration:	Say:
Show S shower curtain ad and say:	Here is an ad for a set of shower and window curtains for a bathroom. The set is on sale for fifty percent off the regular price. How much will it cost now that it is on sale? You may use a paper & pencil if you like.
If S says "$3" within 60 seconds, score 3 and say:	Good. Right.
If S answers incorrectly, say the words in italics followed by the clue. If S is still working after 60 seconds or cannot answer, say only the clue.	*No. That's not quite right.* The curtains usually cost six dollars. Now they are on sale for fifty percent off, or, in other words, one-half off the regular price. What do they cost now?
If S gives correct answer after clue, score 1 and say:	Good. Right.
If S still answers incorrectly, say "No" followed by the explanation. If S is still working or cannot answer 60 seconds after the clue has been given, score 0 and give only the explanation.	*No.* The usual cost is six dollars. Half of six dollars is three dollars. Six dollars minus three dollars leaves three dollars. The sales price is three dollars. Okay?...Good.

Figure 3

Applied Performance Test Item

Comparison of Criterion Validity of Alternate Predictor Formats

1. Sample and Procedures

As concurrently as possible with the administration of predictor instruments, a subsample of the total project sample (students who had taken two forms of the knowledge tests and whose parents had responded to the behavior rating scale) was administered the criterion instrument.

All applied performance tests were administered by the same person, a psychologist with ten years of clinical experience. In all cases, applied performance tests were administered to participants on an individual basis at the school where each participant was enrolled.

These procedures eventually produced 156 completed protocols, which included two knowledge test scores, a behavior rating score, and an applied performance score within all three content areas. Since only two-thirds of this sample took any given form of the knowledge tests, the comparative validity studies have an upward limit of approximately 100 subjects, with the number of subjects available for any given analysis ranging from 71 to 99.

2. Data Analyses

First, basic descriptive statistics characterizing the applied performance tests as instruments were calculated, including means, average difficulty levels, standard deviations, and internal consistency reliability estimates. Each of the five instrument sets were compared along these dimensions. The results of these analyses can be found in Tables 2a - d (pp. 21 to 23)and Table 3 (p. 24) in the Results section.

In order to investigate the narrower research question concerning the differential validity of alternate item formats within the general method of knowledge testing, single-order correlations and simple regressions of item formats upon applied performance were calculated. Intercorrelations among all five instrument were determined, as well as multiple correlations in which knowledge with behavior ratings were regressed upon applied performance. The results of these analyses are presented in Tables 4a - d (pp.25 to 27) in the Results section which follows.

Cross-validation procedures were then carried out in order to estimate the degree of shrinkage to be expected should the obtained simple and multiple correlation regression weights be applied to scores obtained by other samples.

Results

Introduction

Prior to examining the larger question of criterion validity, it is important to demonstrate that the instruments involved possess adequate psychometric properties. Thus, typical descriptive statistics were calculated for all five instruments employed in this investigation, including sets of three types of item formats (true/false, two-option multiple-choice, and three-choice) within the general knowledge testing method

of measurement, as well as the other major predictor method (the behavior rating instrument), and the criterion instrument itself (the applied performance test). The psychometric properties calculated include the mean, standard deviation and difficulty level for each instrument within each of the three domains (Purchasing, Job Search, and Banking) and for the total set of all three domains combined.

In order to answer the research question concerning the differential reliability of various item formats within the general knowledge testing method of measurement, coefficient alpha internal consistency reliabilities were calculated. Coefficient alpha "represents the expected correlation of one test with an hypothetical alternative form containing the same number of items" (Nunnally, 1967, p. 196). It is also the average of all possible split half correlations within a given set of items (Anastasi, 1976).

In order to investigate the questions of validity, intercorrelations of all instruments were computed, and regression analyses were carried out, the latter playing an important role in predictive research frameworks. The rationale for use of regression analysis is clearly explained by Kerlinger and Pedhazur (1973, p. 281): "On the basis of knowledge of one or more independent variables . . ." (in the present case knowledge test alone or a knowledge test plus a behavior rating scale)," . . . the researcher wishes to develop a regression equation to be used for the prediction of a dependent variable, usually some criterion of performance or accomplishment" (in this case, a measure of applied performance).

In order to estimate the degree of shrinkage of the resulting simple and multiple correlations that might be expected when predicting the performance of new samples, cross-validation analyses were performed. For each domain and for the total of all three domains, the sample for which all relevant data were available was randomly split into two subsamples, with the first containing about two-thirds of these subjects and the second including one-third of the sample. The larger of these subsamples was then used as the screening or "keying" sample and the smaller for cross-validation calibration. Using typical cross-validation procedures, a regular regression analysis was performed for the screening sample, calculating R^2 and the regression equation. This regression equation was then applied to the predictor variable(s) of the cross-validation sample, yielding a predicted applied performance score (Y') for each subject. A Pearson r was then calculated between the observed applied performance (criterion) scores (Y) and the predicted criterion scores (Y') in the calibration sample. This Pearson r is analogous to a multiple correlation (Kerlinger, 1973). Finally, an estimate of the amount of shrinkage to be expected when applying the regression equation obtained in the screening sample to future predictions was obtained by calculating the difference between R^2 of the screening sample and R^2 from the calibration sample.

Final Psychometric Characteristics of Instruments

Tables 2a-2d show the basic descriptive statistics that characterize each of the instruments in the study. (From this point, Table a will include data for the total of all three domains, while results for the Purchasing area will be contained in Table b, and Tables c, and d will include information concerning the Job Search and Banking areas respectively). Means for true/false and two-option multiple-choice tests were almost identical (ranging from .68 to .73 across the domains), whereas the mean for the three-option multiple-choice test was slightly lower (ranging from .60 to .63), as would be expected theoretically due to the presence of an additional distractor. Standard deviations for all three knowledge item formats were nearly identical, between 4.00 and 5.00 for each domain and approximately 14.00 for the total of all three domains.

Means on the behavior rating form and applied performance test were very similar to each other within each domain (ranging from the .50's to the 60's), with both somewhat lower than means on the knowledge tests. Variabilities of the behavior rating form and applied performance test were also similar to each other and, proportionally, equivalent to the variabilities of the knowledge tests.

Table 2a

Descriptive Statistics

ALL THREE DOMAINS COMBINED

	SPIB	Multiple Choice-2	Multiple Choice-3	Behavior Rating	Applied Performance
Possible Range	0-99	0-99	0-99	0-183	0-207
N (Sample)	243	261	189	147	156
Mean	70.81	69.30	62.46	108.78	113.63
Difficulty	.72	.70	.63	.59	.55
S.D.	14.07	13.39	13.80	32.74	41.08

Table 2b

Descriptive Statistics

PURCHASING DOMAIN

	SPIB	Multiple Choice-2	Multiple Choice-3	Behavior Rating	Applied Performance
Possible Range	0-36	0-36	0-36	0-63	0-63
N (Sample)	261	287	227	147	158
Mean	26.04	25.78	22.75	38.24	32.44
Difficulty	.72	.72	.63	.61	.52
S.D.	6.06	4.98	5.77	13.27	14.63

Table 2c

Descriptive Statistics

JOB SEARCH DOMAIN

	SPIB	Multiple Choice-2	Multiple Choice-3	Behavior Rating	Applied Performance
Possible Range	0-32	0-32	0-32	0-51	0-60
N (Sample)	261	287	227	147	158
Mean	23.41	21.93	19.79	26.69	33.91
Difficulty	.73	.69	.62	.52	.57
S.D.	4.99	4.75	5.14	11.06	13.11

Table 2d

Descriptive Statistics

BANKING DOMAIN

	SPIB	Multiple Choice-2	Multiple Choice-3	Behavior Rating	Applied Performance
Possible Range	0-31	0-31	0-31	0-69	0-84
N (Sample)	261	287	227	147	158
Mean	21.00	21.30	18.60	43.85	46.41
Difficulty	.68	.69	.60	.64	.55
S.D.	4.79	5.37	5.34	12.56	16.96

Reliability of Alternate Forms

As inspection of Table 3 reveals, estimates of internal consistency reliability, as measured by coefficient alpha, were uniformly high for all five instruments. Sample size is presented in parentheses just below each correlation. Concerning the specific question of differential reliability of item formats within the knowledge testing method, the original true/false item format produced homogeneity indices generally slightly higher than those yielded by the multiple-choice item sets.

Table 3

Reliability Estimates*

	Purchasing	Job Search	Banking	Total
SPIB (True/False)	.83 (261)[a]	.80 (261)	.76 (261)	.92 (243)
Two-Choice Multiple-Choice	.74 (287)	.73 (287)	.82 (287)	.90 (261)
Three-Choice Multiple-Choice	.79 (227)	.76 (227)	.79 (227)	.90 (189)
Behavior Rating Form	.91 (147)	.93 (147)	.91 (147)	.95 (147)
Applied Performance Test	.89 (158)	.88 (158)	.91 (158)	.96 (156)

*coefficient alpha internal consistency reliability

[a]sample size

Validity of Alternate Forms

Tables 4a-4d show zero-order correlations among each of the five instruments, as well as multiple correlations in which knowledge and behavior ratings were regressed upon applied performance.

As might have been expected, the intercorrelations among the three knowledge item formats were moderately high, generally in the .80s for all domains but Job Search, in which intercorrelations ranged from .58 to .68. Perhaps somewhat surprisingly, the correlations between the knowledge tests and applied performance were almost as high, whereas single-order correlations between behavior rating and applied performance were definitely lower, ranging from .47 for the Banking area to .56 for the total of all domains. The correlation between knowledge tests and behavior ratings were lower still, in the .30's and .40's for all content areas. In

each case, the multiple correlation regressing knowledge and behavior rating upon applied performance yielded a value only slightly higher than the correlations between knowledge alone and applied performance.

Table 4a

Correlation Matrix and Multiple Regressions

ALL THREE DOMAINS COMBINED

	SPIB	MC-2	MC-3	BR	AP
SPIB	----				
Multiple Choice-2	.86 (123)*	----			
Multiple Choice-3	.84 (81)	.82 (81)	----		
Behavior Rating	.40 (74)	.49 (99)	.39 (71)	----	
Applied Performance	.78 (74)	.77 (99)	.73 (71)	.56 (140)	----

*Numbers in parentheses indicate sample size for the correlation.

$R = .81$ ($N = 74$), with SPIB and Behavior Rating scores regressed on Applied Performance scores. ($R^2 = .66$)

$R = .79$ ($N = 99$), with 2-choice and Behavior Rating scores regressed on Applied Performance scores. ($R^2 = .62$)

$R = .76$ ($N = 71$), with 3-choice and Behavior Rating scores regressed on Applied Performance scores. ($R^2 = .58$)

Table 4b

Correlation Matrix and Multiple Regressions

PURCHASING DOMAIN

	SPIB	MC-2	MC-3	BR	AP
SPIB	----				
Multiple Choice-2	.75 (125)	----			
Multiple Choice-3	.77 (95)	.71 (96)	----		
Behavior Rating	.37 (77)	.40 (103)	.43 (80)	----	
Applied Performance	.68 (77)	.64 (103)	.68 (80)	.51 (140)	----

R = .72 (N = 77); SPIB + BRF → APT (R^2 = .52)

R = .68 (N = 103); MC-2 + BRF → APT (R^2 = .46)

R = .72 (N = 80); MC-3 + BRF → APT (R^2 = .52)

Table 4c

Correlation Matrix and Multiple Regressions

JOB SEARCH DOMAIN

	SPIB	MC-2	MC-3	BR	AP
SPIB	----				
Multiple Choice-2	.63 (129)	----			
Multiple Choice-3	.68 (97)	.58 (93)	----		
Behavior Rating	.37 (80)	.30 (107)	.28 (80)	----	
Applied Performance	.59 (80)	.60 (107)	.67 (80)	.51 (142)	----

R = .66 (N = 80), SPIB + BRF → APT (R^2 = .43)

R = .68 (N = 107), MC-2 + BRF → APT (R^2 = .47)

R = .70 (N = 80), MC-3 + BRF → APT (R^2 = .49)

Table 4d

Correlation Matrix and Multiple Regressions

BANKING DOMAIN

	SPIB	MC-2	MC-3	BR	AP
SPIB	----				
Multiple Choice-2	.78 (131)	----			
Multiple Choice-3	.62 (85)	.77 (85)	----		
Behavior Rating	.30 (78)	.45 (105)	.29 (76)	----	
Applied Performance	.76 (78)	.72 (105)	.58 (76)	.47 (142)	----

$R = .82$ ($N = 78$), SPIB + BRF → APT ($R^2 = .67$)

$R = .74$ ($N = 105$), MC-2 + BRF → APT ($R^2 = .55$)

$R = .62$ ($N = 76$), MC-3 + BRF → APT ($R^2 = .38$)

Cross-Validation

Concerning the narrower question of differential prediction of performance by different item formats within the method of knowledge testing, the results of cross-validation of these simple regressions show a geater degee of shrinkage within the original true/false item format than within either multiple-choice format (See Tables 5a-5d).

Table 5a

Cross-Validation of Simple Regressions

ALL THREE DOMAINS COMBINED

REGRESSION	Keying Group R	Keying Group R^2	Criterion Group R	Criterion Group R^2	R^2 SHRINKAGE
SPIB → APT	.84 (N=55)	.71	.69 (N=27)	.48	.23
MC-2 → APT	.80 (N=73)	.64	.76 (N=36)	.58	.06
MC-3 → APT	.75	.56	.73	.53	.03

Table 5b

Cross-Validation of Simple Regressions

PURCHASING DOMAIN

REGRESSION	Keying Group R	Keying Group R^2	Criterion Group R	Criterion Group R^2	R^2 SHRINKAGE
SPIB → APT	.78 (N=55)	.61	.61 (N=27)	.37	.24
MC-2 → APT	.71 (N=73)	.50	.54 (N=36)	.29	.21
MC-3 → APT	.63 (N=52)	.40	.65 (N=24)	.42	-.03

Table 5c

Cross-Validation of Simple Regressions

JOB SEARCH DOMAIN

REGRESSION	Keying Group R	Keying Group R^2	Criterion Group R	Criterion Group R^2	R^2 SHRINKAGE
SPIB → APT	.68 (N=55)	.46	.46 (N=27)	.21	.25
MC-2 → APT	.62 (N=73)	.38	.60 (N=36)	.36	.02
MC-3 → APT	.69 (N=52)	.48	.57 (N=24)	.32	.15

Table 5d

Cross-Validations of Simple Regressions

BANKING DOMAIN

REGRESSION	Keying Group R	Keying Group R^2	Criterion Group R	Criterion Group R^2	R^2 SHRINKAGE
SPIB → APT	.76 (N=55)	.58	.74 (N=27)	.55	.03
MC-2 → APT	.73 (N=73)	.53	.73 (N=36)	.53	.00
MC-3 → APT	.60 (N=52)	.36	.63 (N=24)	.40	-.04

In investigating the larger question regarding prediction of performance using different methods of measurement, the earlier reported finding was confirmed, in which it is evident that the addition of the behavior rating instrument to each knowledge test predictor adds little or no new information. In these multiple regression analyses, no consistent pattern of shrinkage appears, with the R^2's of the multiple predictor sets manifesting slightly less shrinkage than did the simple r's of the single knowledge tests alone. The magnitude of such shrinkage and the Banking domain entirely unaffected by shrinkage. See Tables 6a-6d for a summary of the results of the cross-validation of multiple R's and R^2's.

Table 6a

Cross-Validation of Multiple Regressions

ALL THREE DOMAINS COMBINED

MULTIPLE REGRESSIONS	Keying Group R	Keying Group R^2	Criterion Group R	Criterion Group R^2	R^2 SHRINKAGE
SPIB + BRF → APT	.81 (N=74)	.66	.72 (N=24)	.52	.14
MC-2 + BRF → APT	.80 (N=66)	.64	.77 (N=33)	.59	.05
MC-3 + BRF → APT	.76 (N=71)	.58	.73 (N=23)	.53	.04

Table 6b

Cross-Validation of Multiple Regressions

PURCHASING DOMAIN

MULTIPLE REGRESSIONS	Keying Group R	Keying Group R^2	Criterion Group R	Criterion Group R^2	R^2 SHRINKAGE
SPIB + BRF → APT	.72 (N=79)	.52	.61 (N=24)	.37	.15
MC-2 + BRF → APT	.75 (N=66)	.56	.55 (N=33)	.30	.26
MC-3 + BRF → APT	.72 (N=71)	.52	.62 (N=23)	.38	.13

Table 6c

Cross-Validation of Multiple Regressions

JOB SEARCH DOMAIN

MULTIPLE REGRESSIONS	Keying Group R	Keying Group R^2	Criterion Group R	Criterion Group R^2	R^2 SHRINKAGE
SPIB + BRF → APT	.66 (N=80)	.44	.52 (N=24)	.27	.17
MC-2 + BRF → APT	.64 (N=66)	.41	.68 (N=33)	.46	-.05
MC-3 + BRF → APT	.70 (N=31)	.49	.56 (N=23)	.31	.18

Table 6d

Cross-Validation of Multiple Regressions

BANKING DOMAIN

MULTIPLE REGRESSIONS	Keying Group R	Keying Group R^2	Criterion Group R	Criterion Group R^2	R^2 SHRINKAGE
SPIB + BRF →APT	.82 (N=76)	.71	.84 (N=24)	.71	-.03
MC-2 + BRF →APT	.73 (N=66)	.53	.74 (N=33)	.55	-.01
MC-3 + BRF →APT	.62 (N=71)	.38	.66 (N=23)	.44	-.05

Discussion

There is a large user population for assessment instruments and strategies for career education programs for mildly retarded citizens. Educators and vocational rehabilitation professionals consistently voice needs for appropriate measurement tools. Effective programming and placement decisions can result only from precise measurement strategies that are designed to assess specific program components. By exploring test formats that may be appropriate for student evaluation in career education programs, this project has produced results which identify, more clearly than before, guidelines for selection of measurement strategies that can be used appropriately by all who are involved in the education and vocational rehabilitation of mildly retarded citizens.

The specific findings in the current project have a great deal of potential utility. Prior to this project, little progress had been made toward the development of classroom-appropriate applied performance tests for retarded adolescents and young adults in career education programs. This project has resulted in a model for development of such measures, as well as prototype, fully-developed, data-based applied performance tests in three career education domains.

In addition, this project has provided strong empirical support for the utility of the previously developed and widely used Social and Prevocational Information Battery (SPIB). The results demonstrate that a direct testing approach, such as that employed by SPIB, has substantially more validity for predicting applied performance skills than does the behavior rating approach which constitutes much of current assessment practice with retarded persons in career education domains.

Finally, the findings provide direction for future efforts at revision of SPIB or development of new assessment tools for use with retarded individuals. Project results show that retarded persons are able to respond adequately to orally-presented, multiple-choice tests, a test format known to be psychometrically desirable, but previously regarded as inappropriate for retarded persons because of response requirements involving a variety of cognitive skills.

In summary, the project results provide a wide array of data-based guidelines for subsequent efforts to develop and use assessment instruments for placement and programming decisions involving mentally retarded persons.

References

Brolin, D. Programming retarded in career education (Working Paper No. 1). Unpublished manuscript, University of Missouri, September, 1974.

Brolin, D. Preparation of the handicapped for careers. In Proceedings of the Conference on Research Needs Related to Career Education for the Handicapped, January 17-19, 1975, Bureau of Education for the Handicapped, U.S. Office of Education.

Chase, C. I. Measurement for educational evaluation. Reading, Mass.: Addison-Wesley, 1974.

Cronbach, L. J. Further evidence on response sets and test design. Educational and Psychological Measurement, 1950, 10, 3-31.

Ebel, R. L. Expected reliability as a function of choices per item. Educational and Psychological Measurement, 1969, 29, 565-570.

Frisbie, D. A. Multiple-choice versus true-false: A comparison of reliabilities and concurrent validities. Journal of Educational Measurement, 1973, 10, 297-304.

Frisbie, D. A. THe effect of item format on reliability and validity: A study of multiple-choice and true-false achievement tests, Educational and Psychological Measurement, 1974, 34, 885-892.

Gronlund, N. E. Constructing achievement tests. Englewood Cliffs, N. J.: Prentice-Hall, 1968.

Gronlund, N. E. Measurement and evaluation in teaching (2). New York: Macmillan Co., 1971.

Halpern, A. S., Raffeld, P., Irvin, L., & Link, R. Measuring social and prevocational awareness in mildly retarded adolescents. American Journal of Mental Deficiency, 1975, 80, 81-89. (a)

Halpern, A. S., Raffeld, P., Irvin, L., & Link, R. Social and prevocational information battery. Monterey, Ca.: CTB/McGraw-Hill, 19175. (b)

Irvin, L., & Halpern, A. Reliability and validity of the Social and Prevocational Information Battery for mildly retarded individuals. American Journal of Mental Deficiency, 1977, 81 (6), 603-605.

Irvin, L., Halpern, A., & Reynolds, W. Assessing social and prevocational awareness in mildly and moderately retarded individuals. American Journal of Mental Deficiency, 1977, 82 (3), 266-272.

Kerlinger, F. N., & Pedhazur, E. J. Multiple regression in behavioral research. New York: Holt, Rinehart and Winston, Inc., 1973.

Kulman, et.al. The tools of vocational evaluation. Vocatioanl Evaluation Adjustment Bulletin, 1975, 8, 49-64.

Lavisky, S. Invited address. In J. Sanders & T. Sachse (Eds.), Problems and potentials of applied performance testing. Portland, Oregon: Northwest Regional Educational Laboratory, 1975, 33-54.

McClelland D. Testing for competence rather than for 'intelligence'. American Psychologist, 1973, 28, 1-14.

Meyers, C. E., Nihira, K., & Zetlin, A. The measurement of adaptive behavior. In N. R. Ellis (Ed.) Handbook of mental deficiency, in press.

Nunnally, J. C. Educational measurement and evaluation. New York: McGraw-Hill, 1964.

Nunnally, J. C. Psychometric theory. New York: McGraw-Hill, 1967.

Odom, C. L., & Miles, R. W. Oral versus visual presentation of true-false achievement tests in the first course in psychology. Educational and psychological measurement, 1951, 11, 470-477.

Proceedings of the Conference on Research Needs Related to Career Education for the Handicapped. Bureau of Education for the Handicapped, U.S. Office of Education, 1975.

Remmers, H. H., Gage, N. L., & Rummel, J. F. A practical introduction to measurement and evaluation (2). New York: Harper and Row, 1965.

Reynolds, W. M. The effects of multiple-choice formats on test reliability and a search for response styles in mildly mentally retarded adolescents.

Thorndike, R. L., & Hagen, H. Measurement and evaluation in psychology and education (2). New York: Wiley, 1961.

Toward competency: A guide for individualized instruction. Special Education Section, Oregon State Departmetn of Education, 1974.